LEVEL II

(12)

AD A110653

DEPARTMENT OF STATISTICS

The Ohio State University

OSU

COLUMBUS, OHIO

FILE COPY

81 11 25 026

12

OPTIMIZATION IN STATISTICS-AN INTRODUCTION

by

Jagdish S. Rustagi
Department of Statistics
The Ohio State University

Technical Report No. 249

October, 1981

Supported in part by Contract No. N000-14-78-C-0543 (NR 042-403) by Office of Naval Research.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 9	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Optimization in Statistics-An Introduction		5. TYPE OF REPORT & PERIOD COVERED Technical Report
		6. PERFORMING ORG. REPORT NUMBER 249
7. AUTHOR(s) J. S. Rustagi		8. CONTRACT OR GRANT NUMBER(s) N000-14-78-C-0543
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics, Ohio State University, 1958 Neil Ave., Columbus, Ohio 43210		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR 042-403
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research, Department of Navy, Arlington, Virginia 22207		12. REPORT DATE October, 1981
		13. NUMBER OF PAGES 18
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report)
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Optimization, statistical optimization.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Optimization techniques play an important role in various areas of statistics. Examples from Estimation, nonparametric statistics, design of experiments, sample surveys, multivariate statistical inference, information theory and regression analysis are provided.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

1.1 Introduction

In many areas of science, business, industry and government, optimizing techniques are commonly applied to solve routine problems. Topics on optimization have become an important area of study in disciplines such as Operations Research, Chemical Engineering, Electrical Engineering and Economics. Mathematical techniques related to optimization have been developed over the past several hundred years and with the applications of modern computers, these techniques are making an impact in many other areas of science and engineering.

Statistical procedures often require optimization and in a sense one may regard statistics as a subarea of optimization. There are many applications of optimizing methods in the major branches of statistics that the study of optimization becomes an important area for the statistician.

The variety and universality of the use of the optimization techniques can be gauged by a cursory perusal of the contents of the two volumes on Optimizing Methods in Statistics, Rustagi (1971, 1979). The purpose here is to develop a logical introduction to important areas of optimization as they are applied to statistical problems. Several examples are given from statistical areas where optimizing techniques play a major role in their solution. We consider examples from Estimation, Nonparametric Statistics, Design of Regression Experiments, Sample Surveys, Multivariate Statistics, Inference, Information Theory and Regression Analysis to motivate the study of optimization.

The scope of optimizing techniques is fairly extensive. However, we shall put emphasis on those areas of optimization which find frequent applications in statistical problems. The classical techniques

of optimization will be discussed first and numerical methods of optimization will be discussed next. Linear and nonlinear programming methods will be described and variational techniques having connections with dynamic programming and Pontryagin Principle will be discussed later. Applications of other optimizing techniques such as those of Stochastic Approximation will also be included.

1.2 Statistical examples using classical optimizing techniques.

Example: Let $X_1, X_2, \ldots, X_n$ be a random sample from a population having a normal distribution with unknown mean μ and unknown variance σ^2. The estimation of μ and σ^2 by the Method of Maximum Likelihood requires maximizing $L(\mu,\sigma^2)$ where

$$L(\mu,\sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\Sigma(x_i-\mu)^2} \quad . \tag{1.2.1}$$

In this problem, the solution can be obtained by simply equating the partial derivatives of log L to zero and solving the resulting equations to obtain the necessary conditions for an optimum.
Suppose that there is a constraint imposed on the parameter μ, say that μ is always positive. In this case, further attention is to be paid to the process of optimization to obtain the estimate for μ.

Example: Let $p_1, p_2, \ldots, p_k$, $p_i \geq 0$, and $\sum_{i=1}^{k} p_i = 1$ be the probabilities of a trial ending in k possibilities. A sample of n trials, leads to $x_1, x_2, \ldots, x_k$, occurrences of various possibilities. The maximum likelihood estimates of $p_1, p_2, \ldots, p_k$ are obtained by maximizing $L(p_1, \ldots, p_k)$ such that

$$L(p_1, p_2, \ldots, p_k) = p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k} \ , \tag{1.2.2}$$

with constraint, $p_1+p_2+\ldots+p_k = 1$. Usual method of Lagrange's multiplier rule is used to obtain the solution.

Suppose that there are inequality restrictions such as

$$p_1 \leq p_2 \leq \cdots \leq p_k \tag{1.2.3}$$

on the p_i's. In that case the estimates have to use the modern methods of programming.

Example: Consider a normal p-variate population with mean $\underset{\sim}{\mu}$ and covariance matrix $\underset{\sim}{\Sigma}$. Let $\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_N$ be a random sample from the distribution. The logarithm of the likelihood is a constant multiple of $L(\underset{\sim}{\mu},\underset{\sim}{\Sigma})$ where

$$L(\underset{\sim}{\mu},\underset{\sim}{\Sigma}) = -\log|\underset{\sim}{\Sigma}| - \operatorname{tr}\underset{\sim}{\Sigma}^{-1}\underset{\sim}{V} \tag{1.2.4}$$

with $$V = \frac{1}{N}\Sigma'(\underset{\sim}{x}_\alpha-\bar{\underset{\sim}{x}})(\underset{\sim}{x}_\alpha-\bar{\underset{\sim}{x}})' .$$

Again differential calculus provides the maximum likelihood estimates of $\underset{\sim}{\mu}$ and $\underset{\sim}{\Sigma}$. Suppose an additional sample of size M is given on the first $k(k<p)$, of the components. Then the problem becomes more complicated. A recent discussion is due to Anderson and Olkin (1979) for finding Maximum Likelihood Estimates of $\underset{\sim}{\mu}$ and $\underset{\sim}{\Sigma}$. Similar problems arise when some of the components have missing observations.

Example: Bounds of serial correlation coefficient for the time series $x_1,x_2,\ldots$ are needed. The serial correlation coefficient of sth order is defined by

$$r_s = \frac{n}{n-s} \cdot \frac{\sum_{t=1}^{n-s} x_t x_{t+s}}{\sum_{t=1}^{n} x_t^2} . \tag{1.2.5}$$

Consider the upper bound of r_1 for $s = 1$. This is the problem of maximizing r_1 which is equivalent to

$$\max \sum_{t=1}^{n} x_t x_{t+s}$$

subject to the constraint

$$\sum_{t=1}^{n} x_t^2 = \text{constant.}$$

Using Lagrange's method, the solution of the above problem can be obtained Chanda (1962). Contrary to usual belief that correlation coefficient as defined is between -1 and +1, the serial correlation coefficient is defined in (1.2.5) does have higher bounds than 1.

Example: (Constrained regression)

The multiple regression model generally assumes that

$$\underset{\sim}{y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

where $\underset{\sim}{y}$ is nx1 vector, $\underset{\sim}{X}$ is a nxp matrix of known constants, $\underset{\sim}{\beta}$ is a px1 vector of parameters and $\underset{\sim}{\varepsilon}$ is a nx1 random vectors of errors, with means $\underset{\sim}{0}$ and covariance $\sigma^2 \underset{\sim}{I}$. The least squares estimates of $\underset{\sim}{\beta}$ are obtained by

$$\min_{\beta} (\underset{\sim}{y}-\underset{\sim}{X}\underset{\sim}{\beta})'(\underset{\sim}{y}-\underset{\sim}{X}\underset{\sim}{\beta}). \qquad (1.2.6)$$

However, when β is constrained so as to be in a specified set e.g. $\underset{\sim}{\beta} \geq \underset{\sim}{\beta}_0$, we have a constrained optimization problem and in most cases, such an optimization problem requires the use of modern programming methods.

Example: (Optimal allocation in survey sampling)

A large number of problems in survey sampling require optimum allocation of resources since the surveys are constrained by total cost, time or the sampling units. Consider for example the simple case in

cluster sampling where one is interested in determining an optimal size of a cluster which produces the minimum variance of the sample mean for a given cost.

Suppose M be the number of total units to be divided among N clusters of size M_0 each. Let S_W^2 be within-cluster variance and S_B^2 be the between-cluster variance. Let the sample size selected be of size n. Let the corresponding cost C_B associated with a cluster regardless of its size and C_W be the cost associated with each element regardless of cluster size. Then for a fixed cost C, we have

$$C = nC_B + nM_0C_W$$

and we want to minimize the variance of the overall average, $\bar{\bar{y}}$,

$$V(\bar{\bar{y}}) = (1 - \frac{n}{N})\frac{S_W{}^2 + M_0 S_B{}^2}{nM_0} . \quad (1.2.7)$$

The optimal solution turns out to be

$$M_{opt} = \sqrt{\frac{C_B}{C_W}\frac{S_W{}^2}{S_B{}^2}} . \quad (1.2.8)$$

For further details and other problems in sampling using optimization methods, the reader may refer to Jessen (1978) and Cochran (1963).

1.3 Statistical examples using numerical techniques.

Example: Consider the problem of estimation of parameters of the Gamma distribution

$$f(x) = \frac{e^{-x/\beta}x^{\alpha-1}}{\Gamma(\alpha)\beta^{\alpha}} , \qquad x > 0 \quad (1.3.1)$$

$$= 0 \qquad , \qquad \text{elsewhere.}$$

The maximum likelihood estimates are given by equating the partial derivatives of log L with respect to α and β, where

$$\log L = \frac{-\Sigma x_i}{\beta} + (\alpha-1)\Sigma \log x_i$$

$$- n \log \Gamma(\alpha) - n\alpha \log \beta. \tag{1.3.2}$$

The equations are

$$\bar{x} - \hat{\alpha}\beta = 0 \tag{1.3.3}$$

and

$$\Sigma \log x_i - n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - n \log \beta = 0 . \tag{1.3.4}$$

These equations can be solved only through numerical methods. Tables of the diagamma function $\Gamma'(\alpha)/\Gamma(\alpha)$ are available, Pearson and Hartley (1954), to facilitate the solution.

Example: (Survival Analysis)

Suppose the times to death of an individual follow an exponential distribution with parameter λ. The number of r deaths observed at times $t_1, t_2, \ldots, t_r$ are known out of n individuals under study. The study is to be analyzed at time a. The estimate of the parameter λ is obtained by considering the following likelihood

$$L = \lambda^r e^{-\lambda \sum_{i=1}^{r} t_i}(1-e^{-\lambda a})^{n-r} \tag{1.3.5}$$

or $\log L = r\log \lambda - \lambda \sum_{i=1}^{r} t_i + (n-r)\log(1-e^{-\lambda a})$. The likelihood equation for λ is

$$\frac{r}{\lambda} - \Sigma t_i + \frac{ae^{-\lambda a}}{1-e^{-\lambda a}} = 0 \tag{1.3.6}$$

The above equation admits only a numerical solution for λ. The roots of the likelihood equations lead to the maximum likelihood estimates.

Example: (Reliability)

Realistic models in reliability theory and survival analysis require numerical evaluation frequently. Consider the three parameter Weibull

distribution model for the time to failure, for a given individual. The probability density function is given by

$$f_T(t) = \begin{cases} \frac{\beta}{\delta}\left(\frac{t-\mu}{\delta}\right)^{\beta-1} \exp\left[-\left(\frac{t-\mu}{\delta}\right)^{\beta}\right], & \\ & t \geq \mu \\ 0 & t < \mu. \end{cases} \tag{1.3.7}$$

Here β, $\delta > 0$ and $\mu \geq 0$.

Suppose the experiment is conducted over the period $(0,t_0)$ and the times of failures of r_0 individuals out of r on test are given by $t_1, t_2,$

Then the likelihood of the sample is given by

$$L = \frac{r!}{(r-r_0)!\ r_0!} \left(\frac{\beta}{\delta}\right)^{r_0} \prod_{i=1}^{r_0} \left(\frac{t_i-\mu}{\delta}\right)^{\beta-1} .$$

$$\cdot \exp\left(-\sum_{i=0}^{r_0} \left(\frac{t_i-\mu}{\delta}\right)^{\beta}\right) \tag{1.3.8}$$

$$\left\{1 - \exp\left[-\left(\frac{t_0-\mu}{\delta}\right)^{\beta}\right]\right\}^{r-r_0} .$$

The maximum likelihood estimates of μ, β and δ can only be obtained numerically. Several such procedures are available in the literature. Mann, Schafer and Singpurwalla (1974) provide many other models in reliability and survival analysis leading to numerical procedures.

Example: (Curve fitting problem)

Suppose measurements of neutron flux (y) are made in a nuclear reactor at various points (x) and the curve to be fitted is

$$y(x) = A \operatorname{Cos} (Bx+E) + C \operatorname{Cosh} (Dx+E). \tag{1.3.9}$$

At points, x_i, $i=1,2,\ldots,n$, adjusted, y_i, were made and using the assumption that variance of the counts is proportional to its mean. It is of interest to find A, B, C, D and E such that we minimize S with

$$S = \sum_{i=1}^{n} \frac{[y_i - A \text{ Cos } (Bx_i+E) - C \text{ Cosh } (Dx_i+E)]^2}{a_i(A \text{ Cos } (Bx_i+E) - C \text{ Cosh } (Dx_i+E))} \qquad (1.3.10)$$

or to simplify the problem by minimizing S* with

$$S^* = \sum_{i=1}^{n} (a_i y_i)^{-1}[y_i - A \text{ Cos } (Bx_i+E) - C \text{ Cosh } (Dx_i+E)]^2 \ . \qquad (1.3.11)$$

The nonlinear form of the function does not allow us to obtain estimates of A, B, C, D and E in closed form. Hooke and Jeeves (1961) provide a numerical method by "Direct Search" technique for this optimization problem.

Example: (Response surface designs)

Consider the following relationship between the mean μ of a response variable y and the independent variables $\underset{\sim}{x}$ with unknown parameters $\underset{\sim}{\theta}$;

$$\mu = f(\underset{\sim}{x},\underset{\sim}{\theta})$$

where $\underset{\sim}{x} = (x_1,\ldots,x_p)'$ is a p-dimensional "design" variable and $\underset{\sim}{\theta} = (\theta_1,\ldots,\theta_k)$ is a k-dimensional parameter. Since the function f is generally unknown, it is estimated by polynomial of certain degree and then the estimates of $\underset{\sim}{\theta}$ are obtained by some method of estimation.

Let $\hat{y}(\underset{\sim}{x})$ = estimated response at $\underset{\sim}{x}$.

The problem in response surface designs is to find that design, that is, $\underset{\sim}{x}$, such that we minimize

$$J = c \int_X (\hat{y}(\underset{\sim}{x}) - \mu(\underset{\sim}{x}))^2 d\underset{\sim}{x} \qquad (1.3.12)$$

over the class of all $\underset{\sim}{x}$.

Several approaches are available in the literature. Initial impetus was provided by Box and Wilson (1951).

1.4 Statistical applications of mathematical programming.

Mathematical optimization techniques with inequality constraints most often lead to problems of mathematical programming. There is a considerable literature in the application of mathematical programming in statistics, for a recent survey, see Arthanari and Dodge (1981).

Example: (Linear Regression)

A common model of linear regression is the following

$$\underset{\sim}{y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

where y is a vector of n dimensions, $\underset{\sim}{X}$, is a nxp matrix of known constants, $\underset{\sim}{\beta}$ is a p-dimensional vector of unknown parameters and $\underset{\sim}{\varepsilon}$ is an n-vector of residuals. Suppose it is assumed that

$$\underset{\sim}{C}\underset{\sim}{\beta} \geq 0 \tag{1.4.1}$$

where $\underset{\sim}{C}$ is some gxp known matrix. It is proposed to estimate $\underset{\sim}{\beta}$, in general by least squares method so as to minimize

$$(\underset{\sim}{y}-\underset{\sim}{X}\underset{\sim}{\beta})'(\underset{\sim}{y}-\underset{\sim}{X}\underset{\sim}{\beta}) \tag{1.4.2}$$

subject to (1.4.1). The above problem then reduces to the problem of quadratic programming and there are well-known algorithms to deal with such problems. Several authors have contributed to this area, see for example Davis (1978), Judge and Takayama (1966).

Example: (Sampling)

One of the common problems in sample survey is the estimation of the total y, of a finite population given by

$$y = \sum_{i=1}^{T} N_i y_i^* \tag{1.4.3}$$

where $y_1^*, y_2^*, \ldots, y_T^*$ are the possible individual values of items in the population; N_i is the number of units in the population having values y_i^*.

Let n_i be the number of units in the sample with values y_i^* so that under simple random sampling, of size n, we have $n = \Sigma n_i$, the maximum likelihood estimate of y is obtained by maximizing the likelihood

$$L = \prod_{i=1}^{T} \frac{\binom{N_i}{n_i}}{\binom{N}{n}} \tag{1.4.4}$$

with $n = \Sigma n_i$, being the total sample size. The optimization in this case reduces to an integer programming problem, Hartley and Rao (1969) in their paper on A new estimation theory of Sample Surveys. For other problems in survey sampling using mathematical programming, see Rao (1979).

Example: (Design of experiments)

An important class of designs is concerned with factorial experiments. When the number of factors is large, all treatment combinations cannot be used in a block of ordinary size and hence fractional factorial designs have been developed. A recent introduction in the study of fraction factorial is the concept of cost optimality. The problems of finding cost optimal fraction factorials naturally lead to programming problems, Neuhardt and Mount-Campbell (1978).

Example: (Least absolute value estimate in two-way classification).

Consider a two-way classification model

$$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk} \tag{1.4.5}$$

$$i=1,2,\ldots,r,\ j=1,2,\ldots,r,\ k=1,\ldots,n$$

with $\Sigma\alpha_i = \Sigma\beta_j = 0$. y_{ijk} can be regarded as kth observation at the ith level of first factor and jth level of second factor.

We obtain the least absolute value estimates of μ, α_i, β_j by minimizing

$$\sum_i \sum_j \sum_k |y_{ijk} - \mu - \alpha_i - \beta_j|. \tag{1.4.6}$$

This problem is equivalent to the following linear programming problem.

$$\text{Minimize } \sum_i \sum_j \sum_k (d^+_{ijk} + d^-_{ijk}) \tag{1.4.7}$$

subject to

$$\mu + \alpha_i + \beta, + d^+_{ijk} - d^-_{ijk} = y_{ijk} \tag{1.4.8}$$

$$d^+_{ijk} \geq 0$$

$$d^-_{ijk} \geq 0.$$

A large number of other application of programming methods to Least Absolute Value Estimation is found in Gentle (1977).

Example: (Estimation of Markov chain probabilities)

Consider the problem of estimating the transition probabilities $p_{ij}(t)$ of the Markov chain x_t, $t=1,2\ldots,T$, and $i,j=1,2,\ldots r$. Here

$$p_{ij}(t) = Pr\{X_t = s_j | x_{t-1} = s_i\} \tag{1.4.9}$$

where s_i, $i=1,2,\ldots r$ are the finite number of the states of the chain.

$$\text{Here } \sum_i \sum_j p_{ij}(t) = 1 \tag{1.4.10}$$

$$\text{and } 0 \leq p_{ij} \leq t. \tag{1.4.11}$$

Suppose the chain is observed for $N(t)$ independent trials. Let $w_j(t)$ be the proportion of events which fall in jth category. The likelihood of the sample, then, can be obtained as follows

$$L = \prod_{t=1}^{T} \frac{N(t)!}{\prod_m (N(t)w_m(t))!(N(t) - \sum_k N(t)w_k(t))!} \tag{1.4.12}$$

$$\cdot \prod_j (w_i(t-1)p_{ij}(t))^{N(t)w_j(t)}$$

$$\cdot (1 - \sum_k \sum_i w_i(t-1)p_{ij}(t))^{N(t) - \sum_k N(t)w(t)_k}$$

The problem of maximizing the likelihood in (1.4.12) subject to (1.4.10) and (1.4.11) is a nonlinear programming problem. This problem with other manifestations is studied by Lee, Judge and Zellner (1968).

1.5 Variational methods in Statistics.

Classical methods based on calculus of variations have been used extensively in applications, especially for studying physical and mechanical system. Their use in statistics and economics has resulted in new developments of variational methodology. Variational methods are concerned with optimizations of functionals over a class of functions such as minimizing or maximizing integrals of functions over a class of functions subject to certain constraints. In statistics there are many applications which depend very heavily on variational techniques. A recent book on the topic is by Rustagi (1976). We provide a few examples from statistics using classical and modern variational techniques.

Example: (Order statistics)

Suppose $x_1 \leq x_2 \leq \ldots \leq x_n$ is an ordered random sample from a continuous distribution function $F(x)$. The expectation of the largest order statistic X_n is given by

$$L(F) = \int x \, d(F^n(x)). \tag{1.5.1}$$

An important problem in utilizing order statistics is to find upper and lower bounds of $L(F)$ when the mean and variance (say) of the random variable X are given.

Similarly one may want to find the bounds of the expectation of the range, $X_n - X_1$, of the sample. That is,

$$\min(\max) \int x \, d\{1-F^n(x) - (1-F(x))^n\} \tag{1.5.2}$$

subject to certain constraints.

Such problems occur in nonparametric statistical inference and various generalizations have been discussed by Rustagi (1957).

<u>Example: (Mann-Whitney-Wilcoxon statistic)</u>.

Suppose we are interested in the bounds of the variance of Mann-Whitney-Wilcoxon statistic for various applications such as finding confidence intervals for $p = \Pr(X < Y)$. The integral we minimize (maximize) reduces to

$$I(F) = \int (F(x) - kx)^2 dx \tag{1.5.3}$$

subject to the condition

$$\int F(x)dx = 1 - p. \tag{1.5.4}$$

This is a variational problem and has been treated in detail by Rustagi (1961).

<u>Example: (Efficiency of tests)</u>.

Consider a random sample from a population having a continuous distribution function $F(x)$. Suppose we are interested in testing the hypothesis

$$H_0: \quad F(x) = G(x)$$

vs.

$$H_1: \quad G(x) = F(x-\theta)$$

with θ as some location parameter. The relative asymptotic efficiency of Wilcoxon test with respect to t-test (which will be used if F and G were normal distributions) is given by

$$I(f) = \int f^2(x)dx \tag{1.5.5}$$

where $f(x)$ is the corresponding probability density function of X.

A problem of interest in nonparametric inference is to find bounds of $I(f)$ subject to side conditions such as,

$$\int f(x)dx = 1$$

$$\int xf(x)dx = 0.$$

For details, the reader is referred to Hodges and Lehmann (1956).

Example: (Regression designs)

Consider a simple linear regression model,

$$y = \alpha + \beta x + \varepsilon$$

where α, β are the unknown parameters, x is the independent variable and ε is the error with mean 0 and variance σ^2. In regression design of experiments, the investigator is interested in allocating n observations at $x_1,\ldots,x_n$ so as to optimize certain function of the estimated parameters. A common criterion of optimality is the D-optimality where the determinant of the covariance matrix of the estimated factor of parameters is optimized. For example, given a sample of size n the covariance of the $\binom{\hat{\alpha}}{\hat{\beta}}$ is $M\sigma^2$, where

$$M = \begin{pmatrix} \dfrac{\Sigma x_i^2}{n\Sigma(x_i-x)^2} & \dfrac{-\Sigma x_i}{n\Sigma(x_i-\bar{x})^2} \\ \dfrac{-\Sigma x_i}{\Sigma(x_i-\bar{x})^2} & \dfrac{1}{\Sigma(x_i-\bar{x})^2} \end{pmatrix}. \qquad (1.5.6)$$

Assuming that σ^2 is known, the problem of optimal regression design is to find $x_1,x_2,\ldots,x_n$ such that the determinant of M is maximized. There are many other criteria of optimality, a detailed account is available in Federov (1971).

Example: (Robustness)

M-estimates of a location parameter, θ, for a probability density function $f(x-\theta)$, with cumulative distribution function, $F(x-\theta)$, are defined by Huber (1972). A Statistic T_n based on the random sample, $X_1,X_2,\ldots,X_n$ from $f(x-\theta)$ is an M-estimate if it maximizes $\sum_{i=1}^{n} \rho(x_i-T_n)$, for some metric ρ. T_n is given by the equation (1.5.7).

$$\sum_{i=1}^{n} \psi(x_i - T_n) = 0 \tag{1.5.7}$$

with $\psi = \rho'$. Note that if $\rho(x) = -x^2$, we get least squares estimates and if $\rho(x) = -\frac{f(x)'}{f(x)}$, we get maximum likelihood estimates. It has been shown by Huber under fairly general conditions that the asymptotic variance of T_n is $V(\psi,F)$, where

$$V(\psi,F) = \int \left[\frac{\psi(x-T)}{\int \psi'(x-T)F(dx)}\right] F(dx), \tag{1.5.8}$$

with $T_n \to T$ almost surely as $n \to \infty$. The problem is to find an F_0 over the class of functions F which minimizes $V(\psi,F)$. This reduces to a variational problem. Uniqueness and existence of the solutions have been discussed by Huber (1972). Many other problems related to robustness studies leading to the applications of variational methods have been recently discussed by Bickel (1965), Portnoy (1977), and Collins and Portnoy (1979).

Example: (Admissibility)

Let $p_{\underset{\sim}{\theta}}(\underset{\sim}{x})$ be the m-dimensional multivariate normal density of a random vector $\underset{\sim}{x}$. Let $\underset{\sim}{\delta}(\underset{\sim}{x})$ be an estimator of $\underset{\sim}{\theta}$, and let the loss function be

$$L(\underset{\sim}{\theta},\underset{\sim}{\delta}) = (\underset{\sim}{\delta}-\underset{\sim}{\theta})'\underset{\sim}{D}(\underset{\sim}{\delta}-\underset{\sim}{\theta})$$

with $\underset{\sim}{D}$ as a known matrix.

Suppose $G(\underset{\sim}{\theta})$ is the prior distribution function on $\underset{\sim}{\theta}$, $R(\underset{\sim}{\theta},\underset{\sim}{\delta}) = E_{\underset{\sim}{\theta}}\{L(\underset{\sim}{\theta},\delta(\underset{\sim}{x}))\}$, and the Bayes risk is denoted by

$$B(G,\underset{\sim}{\delta}) = \int R(\underset{\sim}{\theta},\underset{\sim}{\delta})G(d\underset{\sim}{\theta}). \tag{1.5.9}$$

Then the Bayes estimator with prior $G(\underset{\sim}{\theta})$, is given by

$$\delta_G(\underset{\sim}{x}) = \frac{\int \underset{\sim}{\theta} p_{\theta}(\underset{\sim}{x})G(d\underset{\sim}{\theta})}{\int p_{\underset{\sim}{\theta}}(\underset{\sim}{x})G(d\underset{\sim}{\theta})} \tag{1.5.10}$$

or

$$\delta_G(\underset{\sim}{x}) = \underset{\sim}{x} + \frac{\Delta g^*(x)}{g^*(\underset{\sim}{x})}, \tag{1.5.11}$$

when $g^*(\underset{\sim}{x}) = \int p_{\underset{\sim}{\theta}}(\underset{\sim}{x})G(d\underset{\sim}{\theta})$ and $\Delta g^*(\underset{\sim}{x})$ denotes the gradient vector of $g^*(\underset{\sim}{x})$.

The sufficient condition for an estimator $\delta_F(\underset{\sim}{x})$ to be admissible is the following:

"There exists non-negative finite Borel measures, G_i $i=1,2,\ldots$ G_i having compact support with $G_i(\{0\}) = 1$, such that

$$B(G_i,\delta_F) - B(G_i,\delta_{G_i}) \to 0 \tag{1.5.12}$$

as $i \to \infty$."

The above condition (1.5.12) reduces to

$$I(g^*,f^*) = \int \left|\left|\Delta\left(\frac{g_i^*(\underset{\sim}{x})}{f_i^*(\underset{\sim}{x})}\right)^{1/2}\right|\right|^2 f^*(\underset{\sim}{x})d\underset{\sim}{x} \ . \tag{1.5.13}$$

Minimizing $I(g^*,f^*)$ answers the problem of admissibility of the estimator $\delta_F(\underset{\sim}{x})$ using techniques of calculus of variations. An elaborate account is in Brown (1971).

<u>Example: (Penalized maximum likelihood estimation)</u>

For various reasons, the estimation of the probability density function $f(x)$ based on a sample $X_1,\ldots,X_n$ is made using a known penalty function $e^{-\Phi(f)}$.

Let the likelihood (penalized) be

$$L(f) = \prod_{i=1}^{n} f(x_i)e^{-\Phi(f)} \ . \tag{1.5.14}$$

The problem of finding penalized maximum likelihood estimates is to find max $L(f)$ subject to constraints,

$$\int f(x)dx = 1$$

$$\text{and} \qquad f(x) \geq 0.$$

This optimization problem reduces to a problem in variational methods. Detailed discussion of this and related problems is given by De Montricher, Tapia and Thompson (1975).

References

Arthanari, T. S. and Dodge, Yadokah. Mathematical Programming in Statistics, John Wiley and Sons, New York, 1981.

Bickel, P. (1965). On some robust estimates of location, Ann. Math. Statist. 36, 847-858.

Box, G.E.P. and Wilson, K. B. (1951). On the experimental attainment of optimum conditions, J. R. Statist. Soc. 13, 1-45.

Brown, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems, Ann. Math. Statist. 42, 885-903.

Cochran, W. G. (1963). Sampling Techniques (2nd Edition), John Wiley and Sons, New York.

Collins, J. R. and Portnoy, S. L. (1981). Maximizing the variance of M-estimators using the generalized method of moment spaces, Ann. Statist. 9, 567-577.

Davis, William W. (1978). Bayesian analysis of the linear model subject to linear inequality constraints, J. Amer. Statist. Assoc. 73, 575-579.

De Montricher, G. F., Tapia, R. A., and Thompson, J. F. (1975). Nonparametric maximum likelihood estimation of probability densities by penalty function methods, Ann. Statist. 3, 1329-1348.

Federov, V. V. (1971). Theory of Optimal Experiments, Academic Press, New York.

Fletcher, R. Practical Methods of Optimization, Vol. I. Unconstrained Optimization, John Wiley and Sons, New York, 1980.

Gentle, J. E. (1977) (Editor). Special issue of Communications in Statistics-Simulation and Computations B 6, 313-438.

Hartley, H. O. and Rao, J.N.K. (1969). A new estimation theory of sample surveys, II, New Developments in Survey Sampling, (Editors, N. L. Johnson and Harry Smith), Wiley-Interscience, New York, 147-169.

Hodges, J. L. and Lehmann, E. L. (1956). The efficiency of some non-parametric competitors of the t-test, Ann. Math. Statist. 27, 324-355.

Hooke, R. and Jeenes, T. A. (1961). Direct search solutions of numerical and statistical problems, J. Assoc. Comput. Mach. 8, 212-219.

Huber, P. (1972). Robust statistics-A review, Ann. Math. Statist. 43, 1042-1067.

Jessen, Raymond, J. (1978). Statistical Survey Techniques, John Wiley and Sons, New York.

Judge, G. G. and Takayama, T. (1966). Inequality restrictions in regression analysis, J. Amer. Statist. Assoc. 61, 166-181.

Kupper, Lawrence L. and Meydrech, Edward F. (1973). A new approach to mean squared error estimation of response surfaces, Biometrika 60, 573-579.

Lee, T. C., Judge, G. G., and Zellner. A. (1958). Maximum likelihood and Bayesian estimation of transition probabilities, J. Amer. Statist. Assoc. 63, 1162-1179.

Mann, N. R., Schaffer, R. E., and Singpurwalla, N. D. (1974). Methods for Statistical Analysis of Reliability and Lifedata, John Wiley and Sons, New York.

Neuhardt, J. and Mount-Campbell, C. A. (1978). Selection of cost-optimal 2^{k-p} fractional factorials, Commun. Statist.-Simula. Computa. B 7, 369-383.

Pearson, K. and Hartley, H. O. (1954). Biometrika Tables for Statisticians, Vol. I and II, Cambridge University Press, Cambridge.

Portnoy, S. L. (1977). Robust estimation in dependent situations, Ann. Statist. 5, 22-43.

Rao, J.N.K. (1979). Optimization in the design of sample surveys, Optimizing Methods in Statistics, (Editor J. S. Rustagi), Academic Press, New York, 419-434.

Rustagi, J. S. (1957). On minimizing and maximizing a certain integral with statistical applications, Ann. Math. Statist. 28, 309-328.

Rustagi, J. S. (1961). On the bounds of variance of Mann-Whitney statistic, Ann. Inst. Statist. Math., 119-126.

Rustagi, J. S. (1976). Variational Methods in Statistics, Academic Press, New York.